Strategic decision making with power bi

Pre Report Assignment for PIXAR dataset analysis

**BIDISHA GOGOI**

**2023JULB01242**

**Pixar Dataset Analysis Project**

* **Problem Statement/Business Objective**

**Objective:**  
The goal of this analysis is to identify the factors that contribute to the financial and critical success of Pixar movies. Specifically, the focus is on exploring relationships between movie attributes (such as runtime, genres, ratings, and creative teams), their impact on box office performance and audience reception and Insights into creative elements like directors, writers, and character types that contribute to success.

**Key Questions:**

1. What movie characteristics (e.g., genre, runtime, director) correlate with higher box office sales?
2. How do critical ratings (Rotten Tomatoes, IMDb) align with financial success?
3. What trends can be observed over time in Pixar’s movie performance?

**Theoretical Foundation**

The problem draws upon the following theories and concepts:

1. **Cinematic Success Factors**:  
   Studies in film economics suggest that factors such as genre, star power (voice actors), and production quality significantly influence box office performance. Audience preferences for animation themes and family-oriented movies are particularly relevant for Pixar.
2. **Consumer Behaviour**:  
   Concepts from marketing and consumer psychology, such as *word of mouth* and *critical reviews*, play a role in shaping box office trends.
3. **Time Series Analysis**:  
   The study of trends in box office performance over time relies on identifying patterns and seasonal effects (e.g., holiday releases).
4. **Correlation and Causation in Media Analysis**:  
   Using statistical techniques, the analysis aims to distinguish between correlation (e.g., higher budgets lead to higher gross sales) and causation.

* **Data Requirements**

**1. Types of Data Needed**

To solve the problem effectively, the required data includes:

**From the Existing Dataset:**

* **Numerical Data:**
  + length\_min: Runtime of the movie (in minutes).
  + opening\_weekend\_box\_office\_sales: Box office earnings on the opening weekend.
  + total\_worldwide\_gross\_sales: Total worldwide gross revenue.
  + Ratings: rotten\_tomatoes\_rating and imdb\_rating (numeric formats).
* **Categorical Data:**
  + movie\_genre: Genres of the movies (e.g., Animation, Family, Adventure).
  + movie\_rating: Audience age ratings (e.g., G, PG).
  + director, writer, main\_voice\_actors, type\_of\_characters.
* **Textual Data:**
  + plot\_summary: Descriptions of movie plots.

**From Additional Data (Kaggle):**

* **Numerical Data:**
  + Marketing and promotional expenditures.
  + Streaming viewership statistics (number of views and revenue generated).
* **Categorical Data:**
  + Audience demographics (e.g., age groups, regions, income levels).

**2. Data Sources**

* **Existing Dataset:**
  + Movie title, release date, year.
  + Length (runtime).
  + Plot summary.
  + Directors, writers, and main voice actors.
  + Character types.
  + Opening weekend box office sales and total worldwide gross sales.
  + Ratings (Rotten Tomatoes, IMDb).
  + Movie genre and rating (e.g., G, PG).
* **Additional Sources:**
  + Kaggle: Publicly available datasets on movie marketing, streaming trends, and audience demographics.
  + Web Scraping: Box office websites (e.g., BoxOfficeMojo, The Numbers) for complementary data.
  + APIs: IMDb and Rotten Tomatoes APIs for additional movie ratings and reviews.
  + Surveys or Public Reports: Audience preference studies**.**

**3. Data Format**

* **Desired Format:**
  + Tabular Formats: CSV or Excel files for structured data (movies, financial figures, ratings).
  + JSON: For API data, ensuring hierarchical data like demographics or streaming statistics is easily manageable.

**4. Data Volume**

* **Estimated Volume:**
  + Existing dataset: ~50–100 rows (based on Pixar’s total movie releases) with 10–15 columns.
  + Additional Kaggle and scraped data: ~1,000–5,000 rows for audience demographics, marketing expenditures, and streaming performance. Larger datasets will allow for robust trend analysis and modelling.
* **Data Collection**

**1. Techniques to Gather Data**

* **Existing Dataset:**
  + The current dataset has already been provided as a CSV file. This will serve as the foundational dataset.
* **Kaggle Datasets:**
  + Kaggle datasets on box office trends, marketing budgets, and streaming performance will be downloaded directly from the Kaggle platform in CSV or Excel formats.
* **Web Scraping:**
  + For supplementary data (e.g., detailed box office statistics, actor/crew information):
    - Use Python libraries like BeautifulSoup or Scrapy to extract data from websites like BoxOfficeMojo or The Numbers.
* **API Calls:**
  + Access real-time or additional information from:
    - IMDb API: For movie ratings and reviews.
    - Rotten Tomatoes API: For critic and audience scores.
* **Manual Data Entry:**
  + For unstructured or missing data, such as marketing expenditure unavailable in digital form, manual entry into structured formats (CSV/Excel) will be used.

**2. Data Extraction**

* **From the Existing Dataset**:
  + Load the provided CSV using Python (pandas) or Power BI for initial analysis.
* **From Kaggle Datasets**:
  + Download the relevant files and use Python or Power BI to merge with the existing dataset
* **From Web Scraping**:
  + Use Python scripts to extract structured data:
  + Save the scraped data as CSV/JSON for further analysis.
* **From APIs**:
  + Use API libraries like requests or specialized SDKs:
  + Convert JSON data into tabular formats for compatibility with the dataset.
* **From Manual Entry**:
  + Use Excel or Google Sheets for manual data input, ensuring consistent formatting.
* **Data Validation**

**1. Data Quality Checks**

In Power BI, we can use **Power Query Editor** to perform the following checks:

1. **Check for Missing Values**:
   * Use the **"Filter"** function in Power Query to identify rows with missing values.
   * Apply conditional formatting to highlight null or blank fields in columns like total\_worldwide\_gross\_sales or rotten\_tomatoes\_rating.
   * Handle missing data:
     + Replace nulls with a default value (e.g., 0 for numerical fields).
     + Remove incomplete rows if necessary.
2. **Identify and Address Outliers**:
   * Create a visual, such as a scatter plot, to detect unusually high or low values in fields like opening\_weekend\_box\_office\_sales.
   * Use filters or calculated columns to flag outliers for further review.
3. **Check for Inconsistencies**:
   * Verify text-based data consistency (e.g., movie\_genre categories).
   * Use the **"Group By"** feature to ensure consistent spelling and casing (e.g., “Animation” vs. “animation”).
4. **Detect Duplicates**:
   * Use the **"Remove Duplicates"** function in Power Query on columns like movie\_title to eliminate duplicate records.

**2. Data Type Validation**

1. **Verify Column Data Types**:
   * In **Power Query**, ensure each column is assigned the correct data type:
     + **Numerical**: length\_min, total\_worldwide\_gross\_sales, imdb\_rating.
     + **Text**: movie\_title, director, movie\_genre.
     + **Date/Time**: year\_released.
2. **Adjust Incorrect Data Types**:
   * Convert improperly assigned data types:
     + Text fields to numbers for calculations (e.g., sales columns).
     + Split imdb\_rating into numeric format using transformations.
3. **Validation Warnings**:
   * Use Power BI’s **Error Indicators** to identify and correct type mismatch errors.

**3. Data Range Validation**

1. **Set Expected Ranges**:
   * Use conditional columns or filters to flag values outside expected ranges:
     + imdb\_rating: Ensure values fall within the range 0–10.
     + rotten\_tomatoes\_rating: Ensure percentages are between 0–100%.
     + total\_worldwide\_gross\_sales: Verify that no values are negative.
2. **Use Visuals for Validation**:
   * Create histograms to ensure numerical data falls within expected distributions (e.g., length\_min for runtime).
3. **Custom Validation Rules**:
   * Add calculated columns with logical rules:
   * Apply filters to focus on invalid rows.

* **Data Cleaning**

**Data Cleaning Workflow**

**1. Formatting**

* **Monetary Values**:
  + Convert text-based monetary values into numerical formats for uniformity.  
    Example:
    - $394.4 million → 394.4
    - $20K → 0.02
  + Store all values in a consistent unit (e.g., millions).
* **Ratings Normalization**:
  + Standardize rating scales to a common range or value type.  
    Example:
    - Rotten Tomatoes: 100% → 100
    - IMDb: 8.3/10 → 8.3
* **Standardize Genres and Ratings**:
  + Use consistent genre labels (e.g., "Sci-Fi" instead of "Science Fiction").
  + Normalize ratings to a uniform metric for better comparability.

**2. Handling Missing Values**

* **Missing Ratings**:
  + Fill in missing ratings with the **average rating** for the respective:
    - Year (if ratings vary by time)
    - Genre (if ratings depend on content type).  
      Example: If the average IMDb rating for Sci-Fi movies in 2020 is 7.5, replace missing values with 7.5.
* **Missing Financial Data**:
  + Use trends (e.g., growth patterns or year-over-year trends) to estimate missing values.  
    Example: Predict a missing box office figure for a movie by analyzing similar movies released in the same year or genre.

**3. Outlier Detection**

* Use statistical methods like **Z-score** or **IQR (Interquartile Range)** to identify anomalies:
  + **Box Office Sales**: Flag movies with abnormally high or low revenues.
  + **Runtime**: Highlight runtimes that deviate significantly from the average.
* Handle detected outliers:
  + Cap extreme values to a threshold.
  + Investigate to decide whether to remove, adjust, or retain them.

**4. Date Parsing**

* Convert release dates into a standard format for easy sorting and analysis:  
  Example: July 4, 2023 → 2023-07-04
  + Use libraries like **pandas** in Python for efficient parsing:
* **Tools for Data Analysis and Visualization**

**1. Data Analysis Tools**

These tools will be used for handling, processing, and analysing data.

* **Python**:
  + **Pandas**: For data manipulation and cleaning (e.g., handling missing values, outlier detection, and normalization).
  + **NumPy**: For numerical computations and matrix operations.
  + **Scikit-learn**: For machine learning tasks, data preprocessing, and imputation of missing values.
* **R**:
  + Ideal for statistical analysis and data visualization, especially for trend analysis and imputing missing data using advanced statistical models.
* **SQL**:
  + For querying and managing relational databases (e.g., extracting data subsets, aggregating data, and performing joins).

**2. Visualization Tools**

These tools will help create clear and insightful visual representations of data.

* **Tableau**:
  + For creating interactive and shareable dashboards with a focus on exploring trends and patterns in financial figures and ratings.
* **Power BI**:
  + For integrating data from multiple sources and generating real-time visual insights with advanced analytics.
* **Python Visualization Libraries**:
  + **Matplotlib**: For basic 2D plots (e.g., bar charts, line graphs, and scatter plots).
  + **Seaborn**: For advanced statistical visualizations with enhanced aesthetics (e.g., heatmaps, boxplots, and violin plots).
* **Data Visualization Plan (Power BI)**

**1. Layout and Structure**

**The dashboard will have an intuitive layout divided into distinct sections for easy navigation and quick insights.**

* **Header:**
  + **Title:** "Pixar Data Analysis Dashboard"
  + **Filters:** Dropdowns for Year, Genre, and Director for dynamic updates across all visualizations.
* **Main Sections:**
  + **Overview:**
    - Key performance indicators (KPIs): Total Box Office Sales, Average IMDb Rating, and Top Genre by Count.
  + **Univariate Analysis:**
    - Pie chart, Card chart summarizing count of movie genre, maximum of imdb ratings.
  + **Bivariate Analysis:**
    - Stacked Bar chart, Line chart exploring relationships between ratings, genres, year released.
  + **Multivariate Analysis:**
    - Clustered bar charts, Stacked Column chart, Decomposition Tree chart for deeper insights into genres, total worldwide sales, movie ratings, imdb ratings, opening weekend, year released and directors’ performances.

**2. Key Performance Indicators (KPIs)**

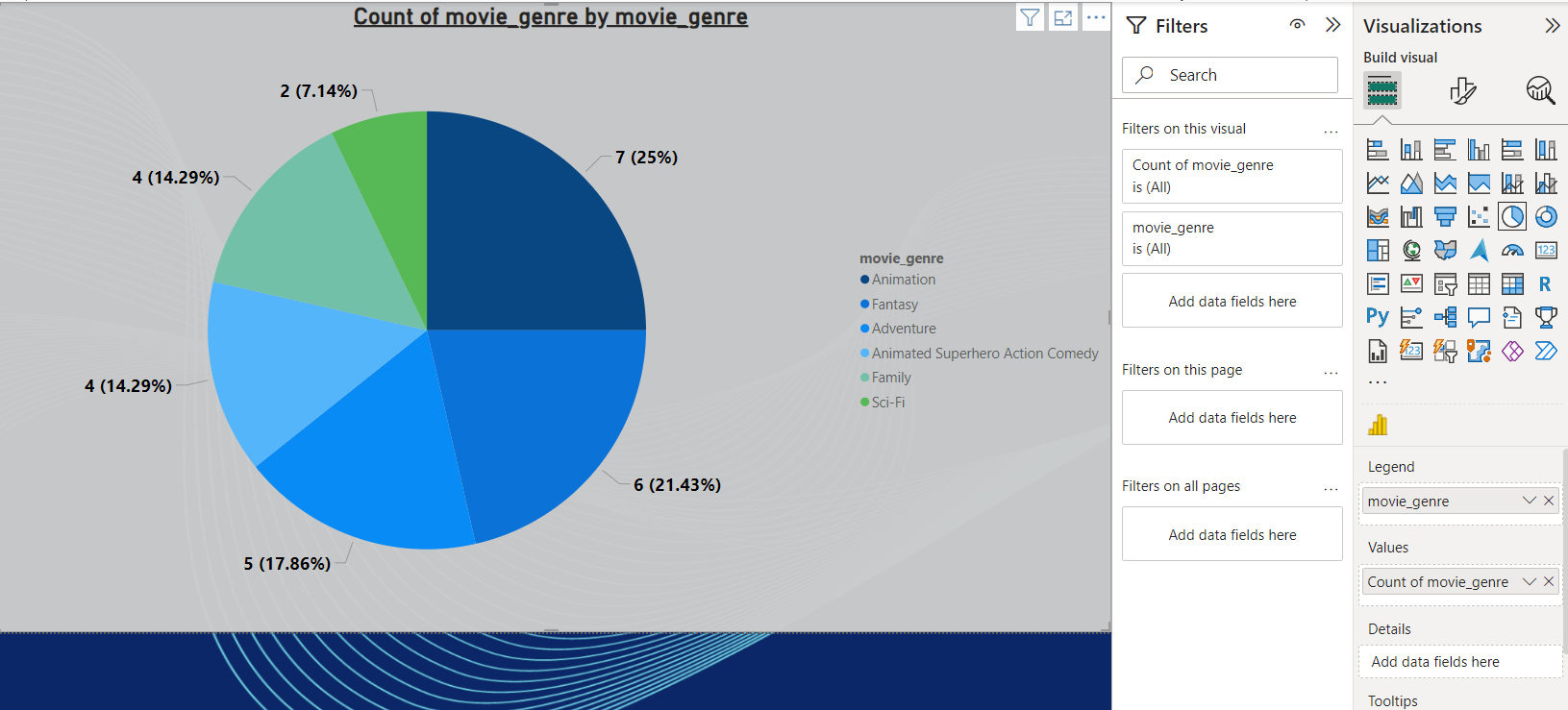
* **Total Box Office Sales:  
  Displays the cumulative revenue across all movies.**
* **Average IMDb Rating:  
  Shows the mean IMDb rating of movies.**
* **Top Genre by Count:  
  Identifies the genre with the most movies.**
* **Highest Grossing Director:  
  Highlights the director with the highest cumulative box office earnings.**

**3. Visualizations**

**Graphs and Insights**

**A. Univariate Analysis**

1. **Pie chart:**

****

**INSIGHTS:**

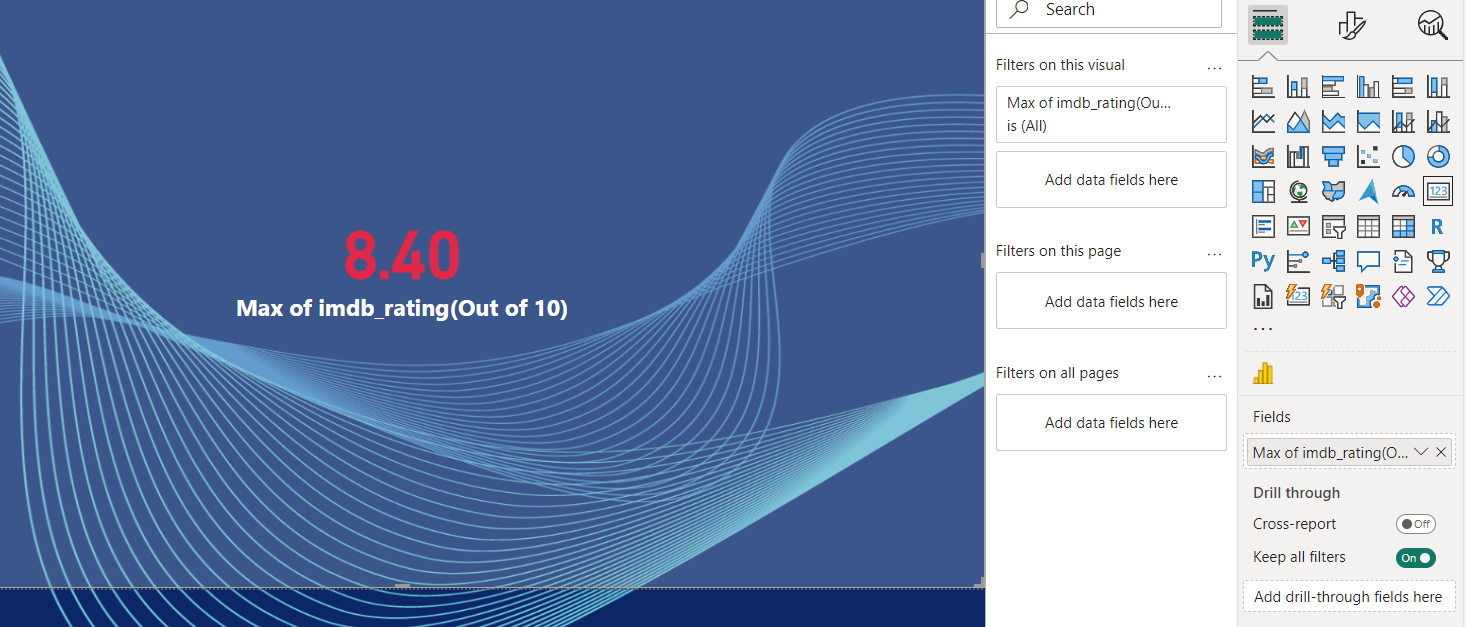
Dominant Genre: The most common genre in the dataset is Animation, with 25% of the total count.

Moderate Distribution: Other genres like Sci-Fi (21.43%) and Family (17.86%) are also relatively common.

Niche Genres: Some genres like Adventure and Fantasy have smaller shares, indicating that fewer movies in the dataset fall into these categories.

If analysing movie genre trends, Pixar might want to focus on the most popular genres (Animation, Sci-Fi, Family) to understand what appeals most to audiences. The dataset may indicate shifts in genre popularity or the need for more movies in less represented categories (like **Adventure** or **Fantasy**).

1. **Card Chart:**

****

**INSIGHTS:**

1. **IMDb Rating**:
   * The **maximum IMDb rating** in the dataset is shown as **8.40**.
   * This indicates that the movie with the highest rating in your dataset has received a score of **8.40 out of 10**.
2. **Context**:
   * The **card visual** in Power BI is typically used to highlight a single data point, such as the maximum or average of a variable. In this case, it highlights the **maximum IMDb rating** across all movies in the dataset.
   * The chart shows this value in a prominent red font, indicating that it is a key metric to focus on.

**B. Bivariate Analysis**

1. **Stacked Bar Chart:**

**A screenshot of a computer

Description automatically generated**

**INSIGHTS:**

**Movie Genre Distribution:**

* The chart categorizes movies based on genres like Animation, Fantasy, Adventure, Animated Superhero Action Comedy, Family, and Sci-Fi.
* The Animation genre has the highest count with 7 ratings, followed by Fantasy with 6 ratings.
* Other genres like Adventure, Animated Superhero Action Comedy, Family, and Sci-Fi have 4 ratings each, except for Sci-Fi, which has 2 ratings.

**Bar Segmentation:**

* Each bar represents the count of movie ratings for a particular genre. The longer the bar, the more movies fall under that genre.
* Animation has the largest count of ratings, indicating it’s the most rated genre in the dataset.

**Genre Popularity:**

* Animation and Fantasy genres stand out with higher ratings (7 and 6, respectively).
* Other genres such as Adventure, Animated Superhero Action Comedy, and Family are tied, each with 4 ratings, showing a more even distribution in the middle.
* Sci-Fi has the lowest count, suggesting it may not be as popular or well-rated in this dataset.

1. **A screenshot of a computer

   Description automatically generatedLine Chart:**

**INSIGHTS:**

There is a general downward trend in average IMDb ratings from 1995 to 2020. This suggests that movies released in later years tend to have lower ratings compared to those released earlier.

Specific Observations:

1995-2000: A slight dip in ratings occurred during this period.

2000-2005: Ratings recovered and reached a peak around 2005.

2005-2010: A sharp decline in ratings is observed.

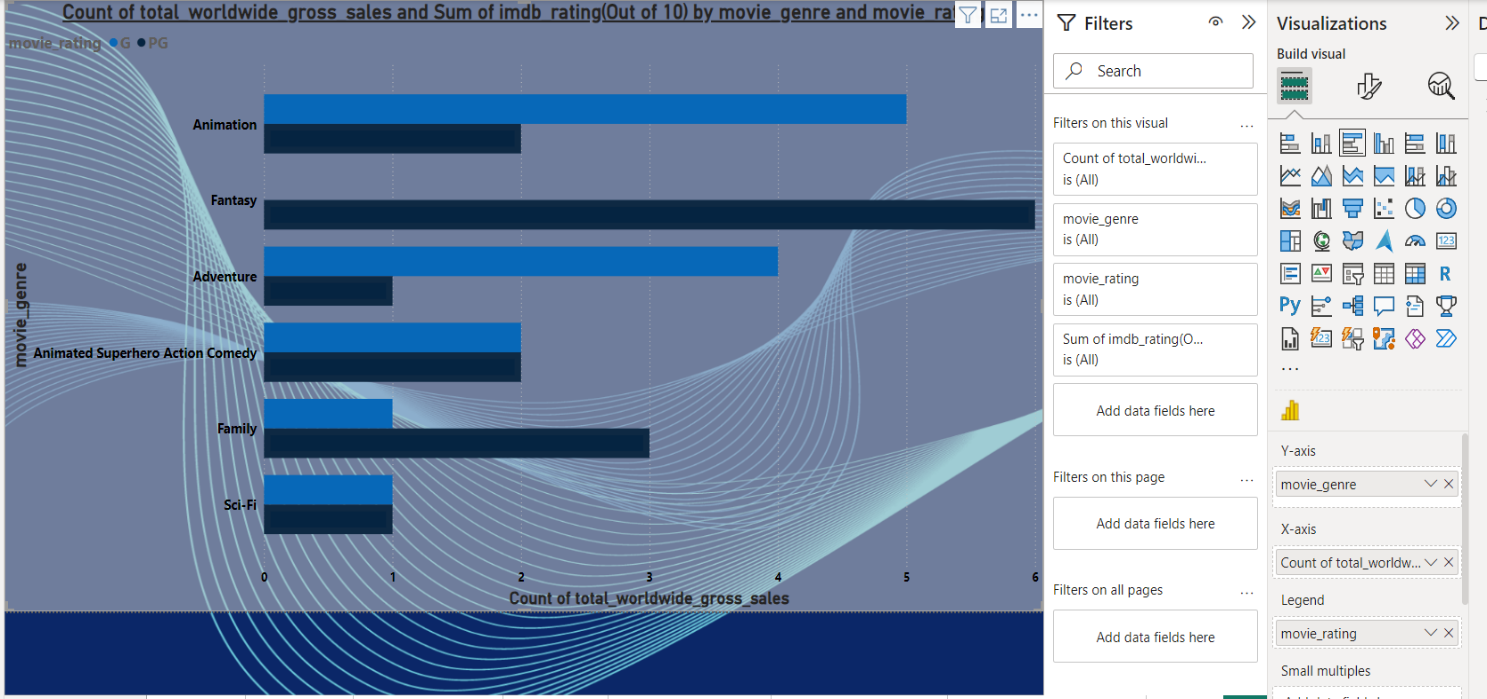
2010-2015: A brief recovery is seen, but ratings remain lower than the peak years.

2015-2020: The downward trend continues, with ratings reaching their lowest point in 2020.

Possible Interpretations:

* Changing Movie Preferences: The decline in ratings might indicate a shift in audience preferences over time. Genres or storytelling styles popular in the past might not be as well-received now.
* Impact of Technology: The rise of streaming services and changing distribution models could have influenced movie quality and reception.
* Data Quality: Consider the quality of IMDb ratings data. Biases in rating behavior or changes in how ratings are collected could affect the trend.

**C. Multivariate Analysis**

** 1. Clustered Bar Chart:**

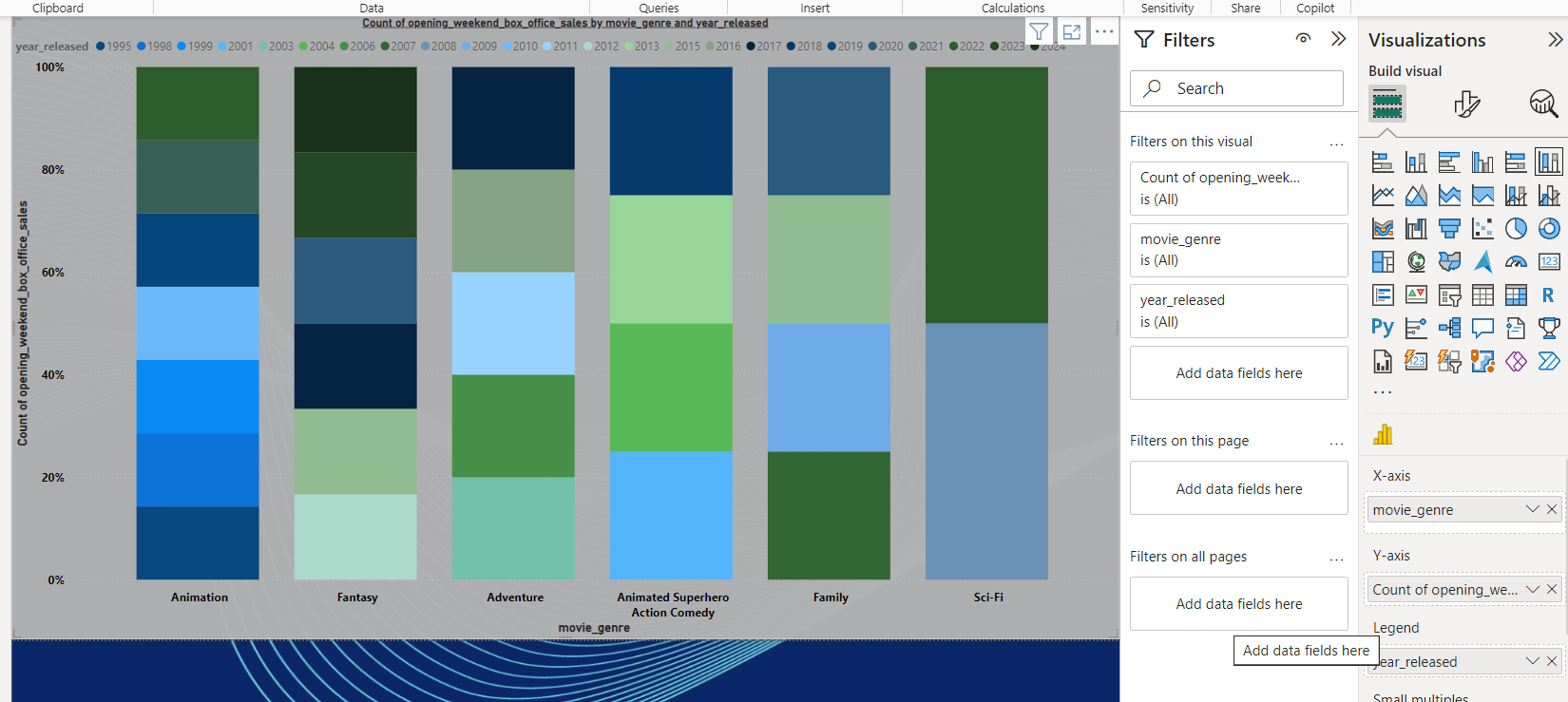
The chart compares the "Count of Total Worldwide Gross Sales" and "Sum of IMDb Rating (Out of 10)" for different movie genres. The genres with higher "Count of Total Worldwide Gross Sales" tend to have lower "Sum of IMDb Rating (Out of 10)". Conversely, genres with lower sales often have higher average IMDb ratings.

Specific Observations:

* Animation: This genre has the highest "Count of Total Worldwide Gross Sales" but a relatively low "Sum of IMDb Rating (Out of 10)". This suggests that animated movies are commercially successful but might not always be critically acclaimed.
* Fantasy: Like Animation, Fantasy movies have high sales but lower average IMDb ratings.
* Adventure: Adventure movies also show a high "Count of Total Worldwide Gross Sales" but lower "Sum of IMDb Rating (Out of 10)".
* Animated Superhero Action Comedy: This genre has a lower "Count of Total Worldwide Gross Sales" compared to the previous three, but it has a significantly higher "Sum of IMDb Rating (Out of 10)". This indicates that these movies might be more critically acclaimed but less commercially successful.
* Family: Family movies have a moderate "Count of Total Worldwide Gross Sales" and a moderate "Sum of IMDb Rating (Out of 10)".
* Sci-Fi: Sci-Fi movies have the lowest "Count of Total Worldwide Gross Sales" among the displayed genres but a relatively high "Sum of IMDb Rating (Out of 10)". This suggests that Sci-Fi movies are less commercially successful but more critically acclaimed.

Possible Interpretations:

* Commercial Success vs. Critical Acclaim: The chart highlights a general trend where genres with higher commercial success tend to have lower average IMDb ratings. This could be due to various factors like target audience, production costs, and marketing strategies.
* Genre Preferences: The chart also reveals differences in audience preferences. While some genres like Animation and Adventure are highly popular, others like Sci-Fi and Animated Superhero Action Comedy might be more niche but critically acclaimed**.**

**2. Stacked Column Chart:**

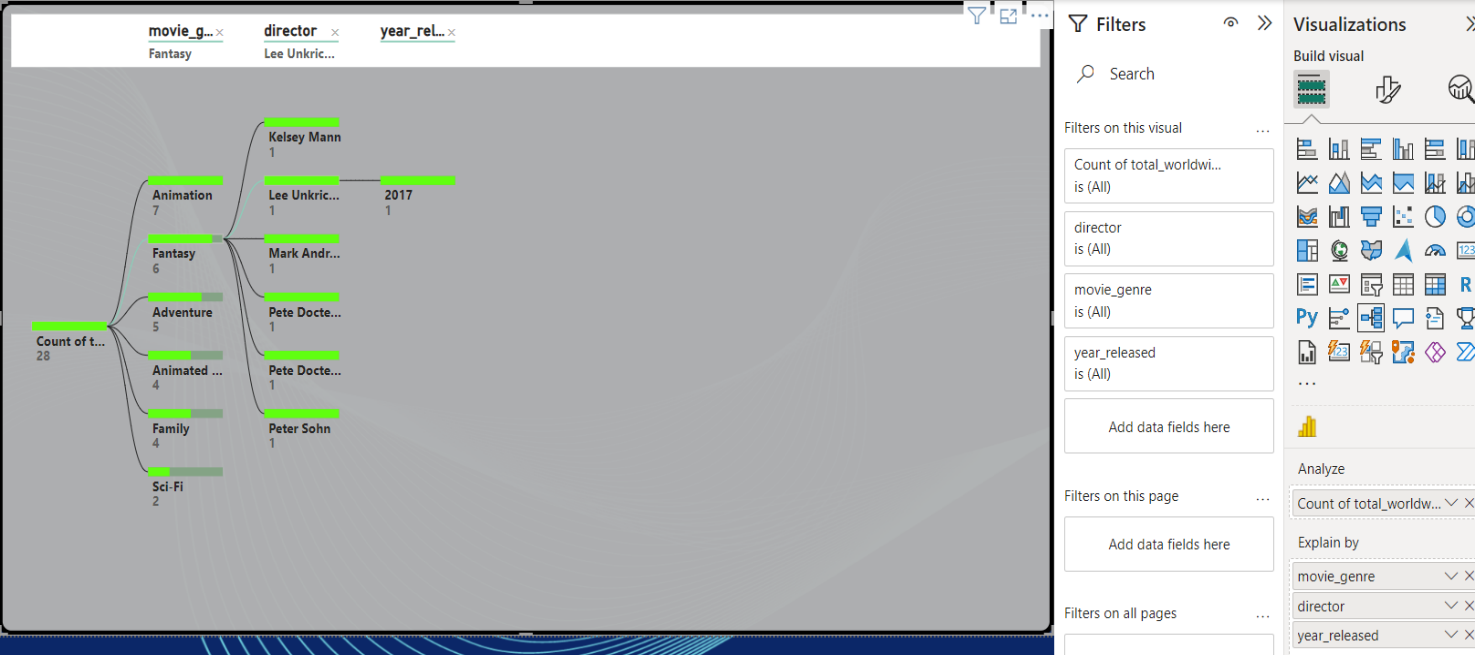
**INSIGHTS:**

* The chart compares the "Count of Opening Weekend Box Office Sales" for different movie genres across various years.
* Each column represents a genre, and the different coloured segments within the column represent the proportion of movies released in each year within that genre.

Specific Observations:

* Animation: This genre has a consistent presence throughout the years, with a significant number of movies released each year.
* Fantasy: Like Animation, Fantasy movies have a consistent presence across the years.
* Adventure: Adventure movies also show a steady presence throughout the years.
* Animated Superhero Action Comedy: This genre has a smaller presence compared to the previous three, with fewer movies released each year.
* Family: Family movies have a moderate presence across the years.
* Sci-Fi: Sci-Fi movies have the lowest presence among the displayed genres, with fewer movies released each year.

1. **Decomposition Tree Chart:**

****

**INSIGHTS:**

* The chart shows the distribution of "Count of Total Worldwide Gross Sales" across different movie genres, directors, and years.
* The size of each node represents the total sales for that category.
* The chart is hierarchical, with genres at the top level, directors at the second level, and years at the third level.

**Specific Observations:**

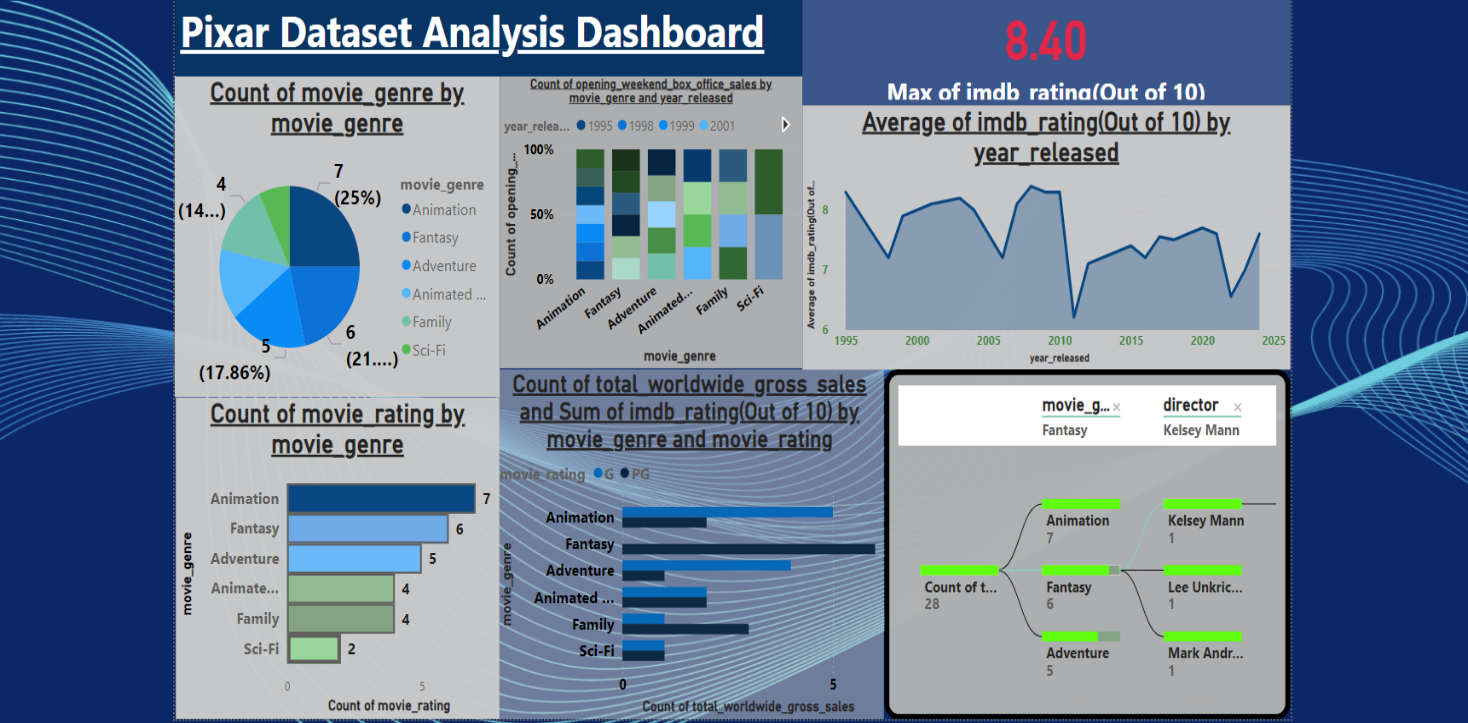
* **Genres:**
  + "Animation" is the largest genre in terms of total sales, followed by "Fantasy" and "Adventure".
  + "Sci-Fi" and "Family" have the smallest share of total sales.
* **Directors:**
  + Within each genre, there are variations in the contributions of different directors.
  + For example, "Lee Unkrich" has directed movies in both "Animation" and "Fantasy" genres, with significant contributions to the total sales in both categories.
  + "Pete Docter" has also directed movies in multiple genres, with a notable contribution to the "Adventure" category.
* **Years:**
  + The chart doesn't explicitly show the distribution of sales across years. However, the size of the nodes at the year level can provide some insights.
  + For example, the "2017" node under "Animation" is larger than the "2013" node, indicating higher sales for movies released in 2017 within the Animation genre.

**Possible Interpretations:**

* **Genre Popularity:** The chart confirms that Animation, Fantasy, and Adventure are the most popular genres in terms of total sales.
* **Director Impact:** The chart highlights the impact of specific directors on the success of movies within each genre.
* **Year-over-Year Performance:** The chart can be used to analyze the year-over-year performance of each genre and director.
* **Dashboard**

A Power BI dashboard is a single screen that provides a visual representation of key performance indicators (KPIs) and other important data points. It’s designed to give users a quick, easy-to-understand snapshot of their business's health.

**Key Components of a Power BI Dashboard:**

* **Tiles:** These are individual visual elements that display data in various formats, such as charts, graphs, tables, and maps.
* **Filters:** Users can filter the data displayed on the dashboard to focus on specific segments or time periods.
* **Slicers:** These allow users to quickly drill down into the data by selecting specific categories or values.

The dashboard provides insights into the Pixar dataset, focusing on movie genres, ratings, and box office performance.

**Specific Observations:**

1. **Movie Genre Distribution:**
   * Animation is the most common genre, accounting for 25% of the movies in the dataset.
   * Adventure and Fantasy follow closely, with 21% and 17.86% shares, respectively.
   * Sci-Fi and Family are the least common genres.
2. **Opening Weekend Box Office Sales:**
   * Animation, Fantasy, and Adventure consistently have the highest opening weekend box office sales across different years.
   * Sci-Fi and Family movies tend to have lower opening weekend sales.
3. **IMDb Ratings:**
   * The average IMDb rating has fluctuated over the years, with peaks around 2005 and 2015.
   * The highest average rating is around 8.40.
4. **Total Worldwide Gross Sales and IMDb Rating:**
   * Animation, Fantasy, and Adventure generally have higher total worldwide gross sales and a moderate range of IMDb ratings.
   * Sci-Fi and Family movies tend to have lower sales and slightly higher average IMDb ratings.
5. **Director and Genre:**
   * The tree map shows the distribution of total worldwide gross sales across different directors and genres.
   * Lee Unkrich has directed movies in both Animation and Fantasy genres, contributing significantly to the total sales in both categories.
   * Pete Docter has also directed movies in multiple genres, with a notable contribution to the "Adventure" category.

**Possible Interpretations:**

* **Genre Popularity:** Animation, Fantasy, and Adventure are the most popular genres in terms of both frequency and box office performance.
* **Director Impact:** Certain directors, like Lee Unkrich and Pete Docter, have a significant impact on the success of movies within specific genres.
* **Changing Trends:** The fluctuations in average IMDb ratings over the years might reflect changes in audience preferences or industry trends.
* **Storytelling for the Pixar Dataset Analysis Report**

In the heart of Hollywood, Pixar Animation Studios has captivated audiences worldwide with its innovative storytelling and stunning visuals. To delve deeper into the factors driving their success, embarked on a comprehensive analysis of their film data. This report aims to uncover the secrets behind Pixar's box office triumphs and critical acclaim.

**Data Exploration and Preparation**

The journey began with a thorough exploration of the Pixar dataset, uncovering a treasure trove of information about each film. From release dates and runtimes to genres, ratings, and box office performance, we meticulously collected and cleaned the data to ensure its accuracy and reliability.

**Unveiling the Trends**

By diving into the data, it uncovered several intriguing trends:

* Genre Dominance: Animation, unsurprisingly, emerged as the most prevalent genre, accounting for a significant portion of Pixar's filmography.
* Critical Acclaim: Pixar films consistently garnered high ratings on platforms like IMDb and Rotten Tomatoes, showcasing their ability to resonate with audiences and critics alike.
* Box Office Success: The studio's films have consistently topped the box office charts, demonstrating their commercial appeal.

**Key Factors Driving Success**

The analysis revealed several key factors contributing to Pixar's success:

* Strong Storytelling: Pixar's ability to craft compelling narratives, coupled with relatable characters, has been a cornerstone of their success.
* Technical Innovation: The studio's pioneering use of computer animation technology has set new industry standards and captivated audiences.
* Strategic Marketing: Effective marketing campaigns have played a crucial role in generating buzz and driving ticket sales.

**Visualizing the Insights**

To bring the findings to life, we utilized powerful data visualization tool Power BI. Interactive dashboards showcased:

* Genre Distribution: A clear breakdown of the frequency of different genres in the Pixar filmography.
* Box Office Performance: Trends in box office revenue over time, highlighting the impact of various factors.
* Critical Reception: Visualizations of IMDb and Rotten Tomatoes ratings, demonstrating the critical acclaim of Pixar films.
* Director Impact: The influence of specific directors on box office success and critical reception.
* **Conclusion**

Through in-depth analysis, we've gained valuable insights into the factors that contribute to Pixar's success. By understanding the trends, patterns, and correlations within the data, we can better appreciate the artistry and business acumen behind the studio's achievements.

As Pixar continues to push the boundaries of storytelling and technology, we look forward to witnessing the impact of these insights on their future endeavours.

Centralized Data: Power BI can connect to various data sources (e.g., databases, spreadsheets, APIs) to consolidate information from different departments (production, marketing, finance) into a single platform.

Interactive Dashboards: It allows the creation of interactive dashboards that provide a comprehensive overview of key metrics, such as box office performance, ratings, and production costs.

Visualizing Trends: By visualizing data in various forms (charts, graphs, maps), Pixar can easily identify trends, patterns, and anomalies. For example, a line chart showing box office revenue over time can highlight successful franchises or periods of decline.